Stacked Omnistereo for Virtual Reality with Six Degrees of Freedom

Jayant Thatte [#], Trisha Lian [#], Brian Wandell [†], Bernd Girod [#]

[#] *Department of Electrical Engineering, Stanford University*
350 Serra Mall, Stanford, CA, USA 94305

[†] *Department of Psychology, Stanford University*
450 Serra Mall #420, Stanford, CA, USA 94305

{jayantt, tlian, wandell, bgirod}@stanford.edu

***Abstract*—Motion parallax is an important cue for depth perception. Rendering it accurately can lead to a more natural and immersive virtual reality (VR) experience. We introduce Stacked Omnistereo, a novel data representation that can render immersive video with six degrees of freedom (DoF). We compare the proposed representation against other depth-based and image-based motion parallax techniques using natural as well as synthetic scenes. We show that the proposed representation can synthesize plausible, view-dependent specular highlights, is compact compared to light fields, and outperforms state-of-the-art VR representations by up to 3 dB when evaluated with 6 DoF head motion.**

I. Introduction

The human visual system infers scene depth by jointly processing several cues. Of these, binocular stereopsis and motion parallax are particularly important within ten meters from the observer [1]. Despite its importance, most natural VR content available today is rendered from a fixed vantage point without accounting for head movements. The resulting mismatch between rendered images and head motion not only breaks realism, but also causes discomfort, nausea [2], and physiological effects similar to car sickness [3].

We propose a novel data representation called Stacked Omnistereo (SOS) that provides convincing 6 DoF rendering of immersive video for viewpoints located within a cylindrical volume. SOS comprises two pairs of vertically stacked stereo panoramas, each augmented with the corresponding depth map. SOS panoramas are designed with an enlarged radius and vertical baseline to ensure that vantage points resulting from a viewer's head motion fall within the cylindrical volume.
The key contributions of our work are:

1) Stacked Omnistereo (SOS) as a novel data representation for 6 DoF VR
2) A systematic comparison of motion parallax fidelity of SOS with other VR representations
3) A method to construct the proposed representation for natural scenes using a 360-degree camera rig [4]

II. Related Work

Omnistereo [5] is widely used in today's VR systems [4], [9], [10]. It provides immersive stereo in horizontal viewing directions from a fixed viewpoint but produces incorrect results with head rotation about a non-vertical axis or any translation. Concentric Mosaics (CMs) have been proposed as an image-based solution for motion parallax [6]. Their main benefit is that novel views can be synthesized by interpolating rays, typically without requiring 3D scene reconstruction. However, because all CM vantage points lie on a single plane, they cannot easily support vertical parallax. CMs take up about 20x more data than omnistereo and are difficult to capture. Rendered views also typically suffer from perspective distortions. These can only be corrected using a depth map, a complication that CMs were developed to avoid.

Depth Augmented Stereo Panoramas (DASPs) were proposed as a depth-based alternative for CMs [7]. DASPs comprise two texture-plus-depth panoramas with an enlarged viewing circle. Within this circle, they can support horizontal head motion almost without artifacts. However, because it is also a planar representation, DASPs must use inpainting to render novel views resulting from vertical translation. While light fields can yield novel views of good fidelity even for highly non-Lambertian scenes, they are associated with a large data volume and their capture can be technically challenging.

Stacked Omnistereo (SOS) builds on the DASP representation and extends it to support 6 DoF. A comparison between SOS and other representations is summarized in Table I.

III. Proposed Representation

A. Construction of Stacked Omnistereo

Stacked Omnistereo (SOS) comprises 2 pairs of texture-plus-depth panoramas on parallel, vertically displaced planes. For simplicity, in this subsection assume that each SOS pixel captures a single incident light ray specified by (1) its position and (2) its orientation and represents RGB color along with a depth value D. SOS geometry is characterized by 2 parameters: radius ρ and height 2λ. The direction and position of the sampled rays is given by[1] $\hat{r} = [1, \theta, \phi]$ and $A = \{\rho\cos\phi, \theta \pm \pi/2, \pm\lambda\}$. The two $\pm$ signs give rise to 4 panoramas. While we use equirectangular projection to map the rays to SOS throughout this paper, it is easily possible to use a different projection, as desired.

[1]Notation: Parentheses, braces, and square brackets denote Cartesian: (x, y, z), polar: $\{\text{radius}, \text{azimuth}, z\}$, and spherical coordinates: $[\text{radius}, \text{azimuth}, \text{elevation}]$ respectively

978-1-5386-0462-5/17/$31.00

TABLE I
COMPARISON WITH OTHER REPRESENTATIONS

Data Representation	DoF: Rotation			DoF: Translation			Needs Depth	No. of Panoramic Images	Specular Highlights
	Yaw	Pitch	Roll	L-R	F-B	U-D			
Stereo Pair	✗	✗	✗	✗	✗	✗	No	2 (images)	NA
Omnistereo [5]	✓	?	✗	✗	✗	✗	No	2	NA
Concentric Mosaics [6]	✓	?	✗	✓	✓	✗	No	$\sim$ 40	Yes
DASPs [7]	✓	✓	?	✓	✓	✗	Yes	2 texture, 2 depth	Limited
Light fields	✓	✓	✓	✓	✓	✓	No	$\sim$ 50-300 [8]	Yes
SOS (Proposed)	✓	✓	✓	✓	✓	✓	Yes	4 texture, 4 depth	Limited

Note: (1) all viewpoints corresponding to a single panorama lie within a disk of radius ρ, called the viewing disk. (2) Except for occlusions, the appearance of each scene point is represented from 4 different directions. The horizontal and vertical baselines are determined by ρ and λ respectively. (3) The cylindrical volume between the two viewing disks is called SOS viewing volume (Fig 1, left). Novel views from viewpoints within this volume can be synthesized with high fidelity using interpolation, usually without inpainting. Hence, it is important to pick SOS height and diameter to be larger than the anticipated range of head motion. (4) Light rays coming in at higher absolute elevation angles get mapped using viewpoints lying on progressively smaller rings with radii given by $\rho \cos\phi$ (Fig 1, right). This allows us to map the whole 3D space, unlike conventional omnistereo [5] which cannot represent the zenith and nadir areas.

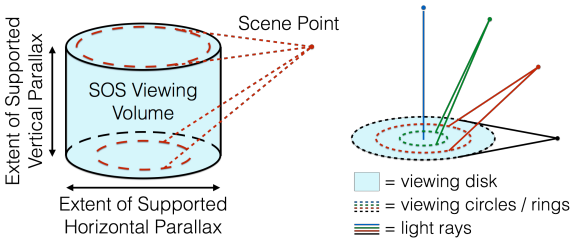

Fig. 1. Left: Each scene point gets recorded using 4 distinct rays. The radius of the dashed circles is proportional to the cosine of the point's elevation angle from the plane. The shaded region denotes SOS viewing volume. Right: Shows a single SOS viewing disk. Mapping regions with increasing elevation angles to smaller rings allows us to map the entire 3D space as shown.

B. Novel Views from Stacked Omnistereo

1) *Forward Depth Warping*: Each SOS depth panorama is mapped to world coordinates knowing A and $\hat{r}$ for each pixel (defined in Sec. III-A). The depth values are then projected onto the target viewport and interpolated using Delaunay triangulation or natural neighbor interpolation [11] to produce one target depth map per SOS panorama.
2) *Backward Texture Lookup*: Each target depth map is used to determine pixel locations in the corresponding texture panorama from which RGB values are fetched using bicubic interpolation. Viewport pixels are converted to world coordinates using depth. A 3D point $P = (x, y, z)$ is mapped to the panoramas using points $A = \{\rho\cos\phi_{CP}, \theta_{CP} \pm \cos^{-1}(\rho/r_{CP}), \pm\lambda\}$, where $[r_{CP}, \theta_{CP}, \phi_{CP}] = \overrightarrow{CP}$. Using equirectangular mapping, $\widehat{AP}$ gives the panorama coordinates. We thus get one target RGBD hypothesis from each SOS panorama.
3) *Hypothesis Merging*: A weighted sum of these hypotheses gives the final viewport. At each pixel (a) smaller depth values take precedence (foreground occludes background), (b) if all depths are roughly equal, then weights are computed based on physical distances between the target and the source light rays. This enables the rendering of view-dependent specular highlights.
4) *Hole-filling*: This is only required for regions that are occluded in all source panoramas, but not from the target viewpoint. These are filled using depth-aware inpainting. In our experiments, less than 0.1% of the pixels needed inpainting for viewpoints in the viewing volume.

C. Camera Rig to Stacked Omnistereo

To support a stereo baseline of 6 cm with 12 cm head motion in any direction for a sedentary viewer, we would need to construct SOS with $12 \times 2 + 6 = 30$ cm diameter ($\sim$5x larger than conventional stereo baseline). Camera rigs [4], [9], [10] are designed to provide stereo with a baseline equal to the human interpupillary distance (6-7 cm). Typically, optical flow is computed on adjacent raw images to synthesize virtual slit cameras along a viewing circle. The slit images are mosaicked into stereo panoramas. This approach is unsuitable for larger viewing radii for two reasons. (1) Optical flow is inherently underconstrained and erroneous flow values lead to jarring artifacts in wide baseline stereo panoramas. (2) Since only adjacent camera pairs are used for 3D reconstruction, the maximum possible viewing radius is severely limited; to simply scale up a camera such as [4], [9] to a viewing circle of 30 cm, we would need a rig diameter of well over a meter.

Fig. 2. A section of the synthesized panoramas is shown. Left three: Using a conventional algorithm [4] that uses only adjacent camera pairs for 3D reconstruction – baselines 6, 12, and 18 cm respectively. Right: SOS panorama with 30 cm diameter using all raw images jointly (proposed approach). While [4] leads to missing regions in panoramas with baselines over 10 cm, the proposed method can handle 5x larger diameter than a typical stereo baseline.

We overcome these challenges as follows. We first rectify and color calibrate all raw images and infer depth from disparity rather than estimating optical flow. Next, we jointly use all the rectified images to construct SOS panoramas following steps similar to Sec. III-B – depth warping, texture lookup,

hypotheses merging, and hole-filling where required. In this context, the target image is an SOS panorama and the sources are the raw rig images. This allows us to generate panoramas with 5x larger diameter without visible distortions (Fig. 2). The depth needed for stitching the raw images from the camera rig is retained as part of the SOS representation. Thus, generating these depth maps is essentially free.

IV. Results

A. Natural Scene

To test the performance on natural scenes, we construct SOS using raw images from a camera rig [4]. We demonstrate that SOS is able to synthesize convincing parallax for head motion trajectories within 30 cm (end-to-end) in any direction (Fig. 3).

Fig. 3. Parallax from 30 cm end-to-end motion. Left: Lateral motion verged on the tree; the chair (solid yellow line) appears to move against the distant trees (dashed red line). Top right: Vertical motion verged on the person; the buildings in the back (dashed red line) appear displaced relative to the person (solid yellow line). Bottom right: Looming motion; the fountain (red circle) stays put while the bushes (yellow square, orange diamond) shift radially.

We also construct Concentric Mosaics (CMs) for these data using $n = 21$ rings with radii $\{0, \sqrt{\rho/n}, \sqrt{2\rho/n}, \ldots, \rho = 15 \text{ cm}\}$ yielding a total of 41 panoramas. Since CMs require several tens of panoramas, constructing the representation is more expensive. Additionally, the rendered images suffer from perspective distortion. SOS corrects them using depth. (Fig.4).

Fig. 4. Left: View rendered using CMs. The building appears skewed and the dome is tilted due to perspective distortion. Right: View synthesized from SOS appears plausible. Quantitative comparison in Sec. IV-B

B. Quantitative Evaluation

Since natural scenes lack ground truth novel views, we use 5 synthetic scenes based on 3D models of natural environments and 1 toy scene with deliberately thin structures and shiny objects. We use a high-quality ray-tracing software called Blender[2]. It solves the light transport equation by Monte Carlo simulation using thousands of rays per pixel yielding photorealistic renderings with specular highlights, reflections, refraction, and indirect illumination.

We compare views from SOS, DASPs, and CMs, using SSIM and PSNR, against ground truth obtained by directly placing a camera at the viewing position in Blender. Using each representation, we synthesize 2,400 viewports with random translations (uniform within the cylindrical viewing volume) and viewing directions (uniform on a sphere). Each view is weighed using the cosine of its elevation angle, similar to [12]. Weighted metrics are denoted with prefix "W-". SOS consistently outperforms both DASPs and CMs for all scenes, averaged across 6 DoF (Table II).

Due to the large number of panoramas needed for CMs ($\sim$40 per scene), we had to limit the representation (and hence viewport) resolution to keep rendering time manageable. We use 2k x 1k panoramas and 333 x 333, 60° viewports. We also compute PSNR on novel views with higher resolution (2k x 2k views from 8k x 4k representation) and 90° field of view (FoV) using SOS and DASPs to test how well the results would scale to typical HMD specifications. Changing resolution does not have a significant impact on PSNR since the representation and the viewports are scaled proportionally. Raising FoV from 60° to 90° leads to $\sim$0.6 dB drop in mean viewport PSNR (250 views across 3 scenes) for both DASPs and SOS. Thus, although the actual numbers show some variation with FoV, the relative comparison still holds.

TABLE II
Comparison with Other Representations

Scene	W-PSNR (dB)			W-SSIM		
	CMs	DASP	SOS	CMs	DASP	SOS
Toy Scene	27.9	36.4	37.4	0.955	0.992	0.994
Bathroom	31.2	34.0	34.6	0.866	0.908	0.914
Kitchen	28.2	32.8	33.7	0.811	0.886	0.898
Seaport	30.3	35.9	36.9	0.842	0.970	0.971
Sunny Rm.	28.2	29.4	30.7	0.933	0.955	0.960
Living Rm.	30.2	34.5	35.1	0.871	0.929	0.933

C. Additional Results

Vertical Translation: SOS outperforms DASPs by up to 3 dB PSNR and 0.01 SSIM depending on the vertical translation and the scene contents. Numerical results are shown on the left in Fig. 5 and example viewports are shown on the right.

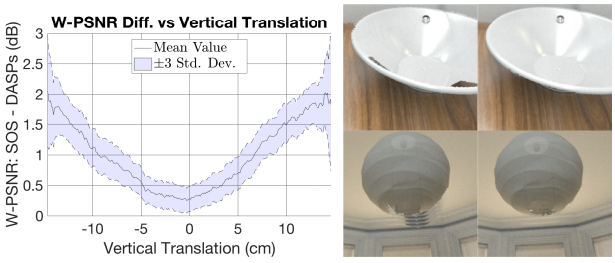

Fig. 5. Left: Mean PSNR gain (SOS over DASPs) as a function of vertical translation. Unlike single-plane DASPs, bi-planar SOS allows it to synthesize vertically translated views usually without needing inpainting. Right: Details of rendered views. In each pair, left image uses DASPs and the right one uses SOS. The wash basin appears broken and the lamp is smeared due to inpainting artifacts. Views are synthesized correctly using SOS.

[2]https://www.blender.org/features/cycles

Perspective Distortion: CMs induce significant perspective distortion in novel views for non-equatorial viewing directions. Since SOS corrects for these using depth, it achieves significantly higher PSNR and SSIM for non-zero elevation angles (Fig. 6). An example for natural scenery was shown in Fig. 4.

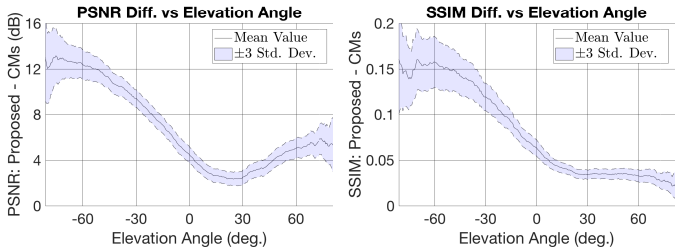

Fig. 6. Mean PSNR gain (left), SSIM gain (right) of SOS over CMs as a function of the viewport elevation angle. SOS uses depth to correct for perspective distortions that exist in views produced using CMs. Smaller gain for positive elevation angles is due to lack of texture in sky/ceiling area.

Specular Highlights: SOS renders inaccurate, but plausible, view-dependent specular highlights (Fig. 8) yielding images that have a very low PSNR, but are visually acceptable (Fig. 7).

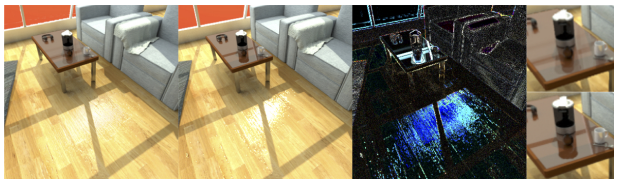

Fig. 7. Left to right: View using SOS, ground truth, absolute difference amplified 10x, and zoomed crop of the table (top: from SOS, bottom: ground truth). SOS produces visually plausible, but incorrect specular highlights leading to PSNR of only 28 dB for the viewport shown.

Fig. 8. SOS novel views from 2 different vantage points. Zoomed crops are shown on right. Notice the parallax as well as changes in specular highlights.

V. Discussion

Limitations: Highly non-Lambertian effects such as reflection and refraction are rendered incorrectly (Fig. 9). This could be addressed in future work along the lines of [13]. Transparent or reflective objects can be handled by modeling scenes as combinations of opaque and additive components thereby allowing multiple depth values per pixel in SOS panoramas.

Fig. 9. Ground truth on left, views from SOS on right. Left pair: Incorrect reflection. Right pair: Ghosting of chair visible through glass pane. Also, the chair with striped back has errors. Background visible through the stripes was occluded in all source panoramas and could not be correctly synthesized.

Other SOS Configurations: Although in this paper we discussed SOS with 2 planes and 2 panoramas per plane, other configurations are possible. The $\{2, 2\}$ configuration is the bare minimum needed to synthesize novel views without inpainting (inside the viewing volume). However, depending on the type of scene and the application, system designers may choose to add more planes or more concentric panoramas per plane, which would lead to better non-Lambertian reconstruction at the cost of larger representation size.

VI. Conclusions

We propose a novel immersive video representation called Stacked Omnistereo (SOS) that can render VR experiences with full 6 DoF head motion in a limited viewing volume. Unlike prior VR representations such as DASPs [7], SOS has a bi-planar structure, which allows it to render motion parallax without inpainting not only for horizontal movements, but also vertical. It outperforms DASPs by a PSNR margin of up to 3 dB and a SSIM improvement of up to 0.01 for vertical translation, averaged across all test scenes. Additionally, SOS uses depth to correct perspective distortions in synthesized views. As a result, in comparison with Concentric Mosaics [6], it achieves a PSNR gain of up to 12 dB and a SSIM gain of up to 0.15 for non-equatorial viewing directions when averaged over all test scenes. SOS is also significantly more compact than light fields. We show that scenes that do not contain highly non-Lambertian effects such as reflection, refraction, or transparency, can be rendered using SOS to provide convincing motion parallax and view-dependent specular highlights.

References

[1] J. Cutting and P. Vishton, *Perceiving layout and knowing distances: the interaction, relative potency, and contextual use of different information about depth*. Cambridge, MA, USA: Academic Press, 1995.

[2] H. Ujike, T. Yokoi, and S. Saida, "Effects of virtual body motion on visually-induced motion sickness," in *Int. Conf. of IEEE Engineering in Medicine and Biology Society*, vol. 1, 2004, pp. 2399–2402.

[3] C. T. Lin, S. W. Chuang, Y. C. Chen, L. W. Ko, S. F. Liang, and T. P. Jung, "EEG effects of motion sickness induced in a dynamic virtual reality environment," in *Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society*, 2007, pp. 3872–3875.

[4] (2016) Facebook Surround360. Accessed Sep. 13, 2017. [Online]. Available: https://facebook360.fb.com/facebook-surround-360

[5] S. Peleg, M. Ben-Ezra, and Y. Pritch, "Omnistereo: panoramic stereo imaging," *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 23, no. 3, pp. 279–290, 2001.

[6] H.-Y. Shum and L.-W. He, "Rendering with concentric mosaics," in *Proc. of the 26th Annual Conf. on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99, 1999, pp. 299–306.

[7] J. Thatte, J.-B. Boin, H. Lakshman, and B. Girod, "Depth augmented stereo panorama for cinematic virtual reality with head-motion parallax," in *IEEE Int. Conf. on Multimedia & Expo (ICME)*, 2016, pp. 1–6.

[8] S. C. G. Laboratory. The (new) Stanford light field archive. Accessed Sep. 13, 2017. [Online]. Available: http://lightfield.stanford.edu/lfs.html

[9] R. Anderson, D. Gallup, J. T. Barron, J. Kontkanen, N. Snavely, C. Hernández, S. Agarwal, and S. M. Seitz, "Jump: Virtual reality video," *ACM Trans. Graph.*, vol. 35, no. 6, pp. 1–13, Nov. 2016.

[10] Cinematic VR field guide - Jaunt. Accessed Sep. 13, 2017. [Online]. Available: https://www.jauntvr.com/cdn/uploads/jaunt-vr-field-guide.pdf

[11] R. Sibson, "A brief description of natural neighbor interpolation," V. Barnett, Ed. Interpreting Multivariate Data: J. Wiley & Sons, 1981.

[12] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE Int. Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[13] S. N. Sinha, J. Kopf, M. Goesele, D. Scharstein, and R. Szeliski, "Image-based rendering for scenes with reflections," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Jul. 2012.